# Stochastic Gradient Descent

## STAT 672 Project

Tom Wallace

George Mason University

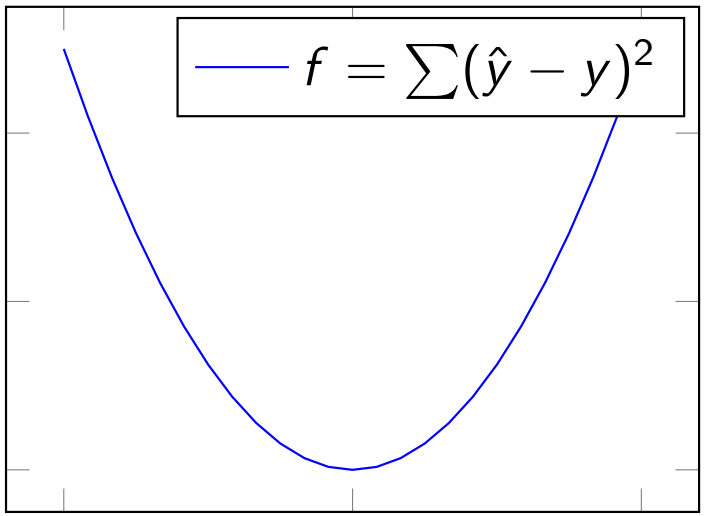
Spring 2018

# Optimization is everywhere, and sometimes is easy

Many statistical procedures involve minimizing or maximizing some function applied to data

In **parametric** statistics, we often make assumptions that make this optimization "nice":

- Example: in OLS, we do not need to check every possible value of $\hat{\beta}$ to see if it minimizes the loss function, we (typically) can just evaluate $(X'X)^{-1}X'Y$



# Other times, optimization is not so easy

Suppose that we have a typical supervised classification problem:

- Non-parametric: no assumptions about distribution of data
- Feature vector $\mathbf{X}_i$, label $Y_i$
- Want to find best prediction function $f^*$ from class $\mathcal{F}$
- $f$ has parameter vector $\boldsymbol{\theta}$
- Optimization: pick values for $\boldsymbol{\theta}$ that minimize empirical risk according to some loss function $L$
- Assume $L$ is convex (if not, problem is *much* harder)

Our lack of assumptions requires a different approach to optimization

- Cannot analytically identify stationary point
- Need to numerically search for it

# Gradient descent is a numerical approach to optimization

Reminder: gradient is multidimensional version of 1st derivative

- $\nabla f(\mathbf{X}) = (\frac{\partial f}{x_1}, \frac{\partial f}{x_2} ...)$
- So if the gradient is (approximately) 0 at some point, that point is (approximately) a stationary point
- Since we know the loss function is convex, that stationary point must be the global minimum

Give basic idea: take guess, step, evaluate, stop once below epsilon

# Batch gradient descent, animated

# Batch gradient descent is computationally expensive

Elaborate

# Stochastic gradient descent (SGD) is more efficient

Overview of how it works

# SGD, visualized

# Choice of hyper-parameters is important

Reminder of what they are

Step size from Bottou

Learning parameter

# SGD has nice properties for high-dimensional data

Theoretical explanation

Empirical performance

# Applications of SGD

If a Silicon Valley press release uses any of the following phrases...

- "Neural networks"
- "Machine learning"
- "AI"

...SGD probably is involved. Example: Google's `AlphaGo` program.